Top 2000 Spotify Songs: Visualization and Analysis

Penelope Prochnow, Joshua King, Kaitlyn Johnson

DSCI – 4013

December 6, 2023

I. Introduction

Embark on a melodic journey through time with the Spotify 2000 song dataset, a curated collection of 2000 songs spanning over six decades from 1956 to 2019. Crafted by Sumat Singh and available on Kaggle, this dataset opens a gateway to the heart of musical expression. Each song, a unique canvas of artistry, is meticulously characterized by its Title, Artist, Genre, Year, and an array of musical attributes. From the pulsating Beats Per Minute (BPM) to the vibrant danceability and the intricate acoustic nuances, the dataset unravels the multifaceted tapestry of musical intricacies. Join us as we explore the rhythmic patterns, dynamic energy, and emotive valence that define each composition. Delve into the sonic landscapes, feel the connection between performers and audiences through liveness, and witness the temporal evolution of music through the Length (Duration) metric.

This dataset is not just a collection of songs; it's a portal to deciphering the language of music, asking questions about the correlation between musical dimensions and popularity, the impact of BPM on chart rankings, and the authentic connection between live performances and audience appeal. Through the compilation of these various measures, a vast arrangement of metrics can be observed leading to the following research questions:

1) Through the past six decades, has there been a profound fluctuation of characteristic variables? Are there distinguishable years in which certain aspects or genres were more prominent?

2) To what extent do characteristic variables of music tend to show predictability of genre or popularity?

3) Are there any artists that have contributed a discernable impact to the top 2000 songs on Spotify?

To ensure precision and relevance, the dataset was judiciously refined by focusing on the 10 most populated genres within the Top 2000 songs. A further refinement involved filtering genres with less than 50 artists, thereby optimizing the dataset for insightful analysis.

II. Visualization and Graphing Methods

This carefully crafted interactive dashboard goes beyond typical data visualizations. It is a thoughtful blend of visual design theory and psychology, offering users a guided journey through the intricate world of music analytics.

At the core is a radial Sankey chart (Figure 1), strategically placed as a viewer's hub – a central point for orientation and focus. This dynamic nexus acts as a go-to point, ensuring users always have a clear sense of direction amid the abundance of data. Designed for simplicity and ease of use, it serves as a starting point for an analytical journey, where selective tools allow for precision in filtering by decades and genres.

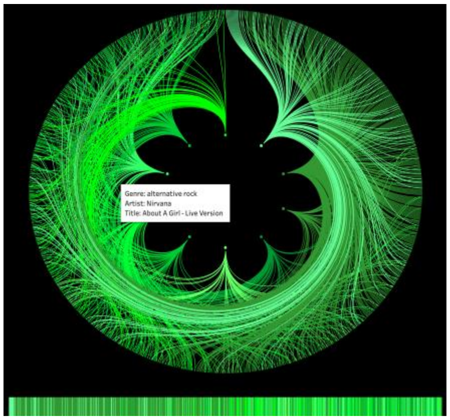

Figure 1. Sankey Diagram, Tooltip Viewable

The genre bar below the Sankey chart, referred to as the edge bar, is an innovative visualization inspired by the seminal insights of Cleveland and McGill [1]. Drawing from their observations on human error in perceiving length within visualizations, particularly around radial distances, the edge bar serves as a refined solution. This innovative element not only mitigates the challenges posed by circular presentations but also serves as an elegant chronological narrative. Genres are presented in a left-to-right sequence, reflecting the order of their release, be it within the selected decade or across the entire dataset. The interactive feature shines through as the edge bar dynamically highlights genres based on the user's selection from the genre chart.

The coordinated interplay of visuals serves functionality to increase understanding of the various encoded variables. Filter by decade and the entire dashboard seamlessly adapts to showcase analytics from that specific time, while the Sankey chart elegantly connects the nuanced snapshot being examined. Select a genre, and the emphasis shifts, highlighting the insights that revolve around a specific musical aspect.

In the realm of psychology, this design is a conscious effort to prevent users from feeling overwhelmed by the wealth of data. The concept of having a hub provides a psychological safety net for the navigator, ensuring a sense of rootedness and clarity amidst the analytical journey. As users explore the upper and lower regions, encountering analytical wonders, the hub remains a reassuring constant.

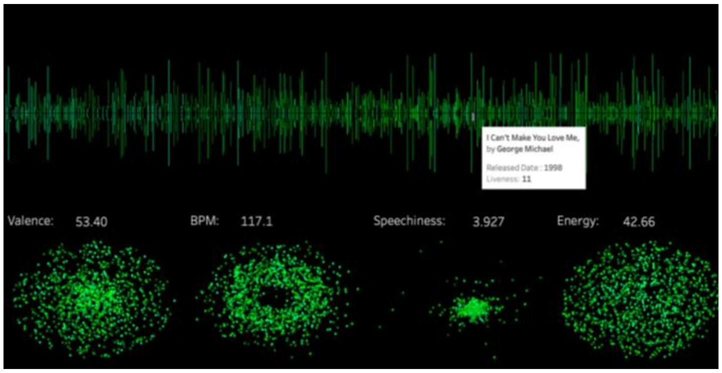

Figure 2. Soundwave sorted by Year accompanying select Radial Density Plots, Tooltip Viewable

The soundwave-inspired bar chart (Figure 2) at the top, a visual masterpiece, is not just an aesthetic choice – it's a focal point representing the liveliness of songs. Below, the radial scatter plots (lower section, Figure 2), designed with analytical precision in mind, dissect musical metrics with finesse. This section serves as a visual encyclopedia, offering detailed insights into the intricacies of musical composition. Without any filters selected, these plots present the numerical average for all values in the dataset, but once a filter is selected, the averages along with the plots themselves transform to reflect the chosen filter. Overarching conclusions can be interpreted from clusters within each plot or specific insight can be derived from the accompanying filters.

The hub's interaction allows for nuanced metrics per decade, genre, or even individual songs.

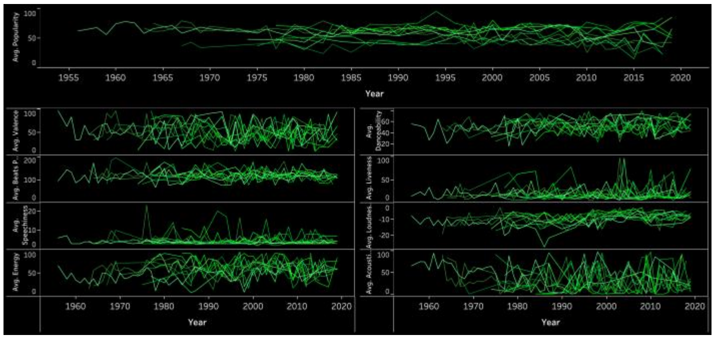

Figure 3. Time Series Line Charts, showing Popularity and Characteristic Variables

In the comparative analytics section below, the time series line charts (Figure 3) for each metric unfold as a crucial exploration into the evolution of music. Delving into each genre with precision, these charts offer a nuanced view of how metrics have evolved, revealing hidden trends that transcend genre boundaries. With a granular perspective, these charts become a chronological window into the dynamic pulse of the Top 2000 Spotify charts.

Legends, crafted with the navigator in mind, strategically minimize strain and confusion. Their black-and-white simplicity aligns with the changing dashboard dynamics, while their proximity to visuals minimizes the need for constant scrolling.

The final crowning element of the dashboard, the popularity chart serves as the pièce de résistance, offering an additional layer of insight. Meticulously positioned to elevate the user experience, this dynamic chart reacts responsively to genre and decade selections, presenting an ordered list of songs based on a discerning popularity metric. Each entry on the list gracefully unfolds with the accompanying genre and song name, delivering a comprehensive snapshot of the musical landscape.

The interactive prowess of the popularity chart extends beyond passive observation. Selecting a song from the list triggers a harmonious synchronization with the remaining visualizations, seamlessly highlighting the chosen composition throughout the dashboard. This not only facilitates a deeper dive into the selected song's contextual relevance but also reinforces the interconnected narrative woven across all elements.

III. Methods: Mathematical Aspect

The transformation of the dataset follows from various metrics that had to be adjusted through mathematical and computational calculations to properly translate to accurate and meaningful conclusive visualizations. In the radial scatterplots, the energy, danceability, loudness, liveness, valence, acousticness, speechiness, and beats per minute measures were normalized to fall within a standard distribution between 0 and 1 to be cross-comparable among all listed characteristic variables. Once normalized, various trigonometric calculations were implemented to create a spread in the data that would adequately show clusters of each characteristic, ultimately providing more intensive insight into the overall quantitative values of each metric.

$$\text{X-axis Coordinates} = \mu \cdot \cos\left(\frac{\pi}{2} - \frac{2\pi}{(n-1)(ID-1)}\right)$$

$$\text{Y-axis coordinates} = \mu \cdot \sin\left(\frac{\pi}{2} - \frac{2\pi}{(n-1)(ID-1)}\right)$$

The above calculations are used to transform the scatter plot of each metric to polar coordinates in the circular scatter plots. These coordinate calculations have a scaling factor of the average measure (µ) for each respective metric that is studied, this variable varies depending on which measure is being studied; to transform the points to be properly represented as polar coordinates, the cosine and sine parts of the equations shift the points utilizing count (n) and index (ID) to properly spread the points around the circular plot. This results in the minimum value for each of the metrics to be centered in each graph and the maximum measures to lie on the outer edges with each song arranged alphabetically clockwise around each plot.

Below are the calculations for the radial Sankey chart. In the carefully curated table below you can see the mathematical formulae for the curvature of paths, distances, coordinates, and classifications.

A. *Mathematical Path from Inner Circle to Outer:*

If categorized as a path:

$$\frac{(Outer\ Circle\ Rank - Modified\ Rank\ \cdot T) + Modified\ Rank}{Count\ of\ Outer\ Circle}$$

If categorized as Inner Circle point:

$$\frac{Inner\ Circle\ Rank}{Count\ of\ Inner\ Circle}$$

If categorized as the Outer Circle point:

$$\frac{Outer\ Circle\ Rank}{Count\ of\ Inner\ Circle}$$

B. *Distances:*

For the line between inside and outside values:

$$\text{Distance} = 2.5 - \frac{2.5}{1+e^{Logodds}}$$

For Inner Circle spacing:

$$\text{Distance} = 2.5$$

For Outer Circle spacing:

$$\text{Distance} = 2.3$$

C. *Logodd, Curvature of Path:*

For each Inner and Outer Circle point as well as the path between, where T signifies the categorization of whether the point is a source (inner circle point), edge (outer circle point) or the path between:

$$\log\left(\frac{T}{1-T}\right)$$

D. *X Values:*

For the lines between Inner Circle and Outer circle points, and spacing of Inner Circle points:

$$(DISTANCE\ +\ 1) \cdot \cos(PATH \cdot 2\pi)$$

E. *Y Values:*

The opposing horizontal spacing, to neutralize vertical spacing causing the array of points to properly splay across the Sankey Diagram:

$$(DISTANCE\ +\ 1) \cdot \sin(PATH \cdot 2\pi)$$

IV. Usability Analysis

The dashboard strategically arranges visual elements for a seamless journey from circular to linear shapes as the viewer scrolls. Beginning with circular radial scatter plots at the top, followed by a transitional radial Sankey chart in the middle, and concluding with linear time series line charts at the bottom, this intentional progression enhances both visual appeal and organizational clarity. This meticulous design, inspired by psychology, caters to innate human preferences, creating a visually engaging and analytically seamless experience. In essence, this meticulously crafted dashboard serves as a testament to the thoughtful planning behind its creation – a seamless fusion of visual aesthetics, interactive functionality, and psychological considerations. It not only addresses quantitative inquiries, such as artist and song counts, but also caters to qualitative curiosities, enabling users to discern genre trends, explore iconic songs from each decade, and track popularity dynamics over time. With its harmonious integration of visual elements and interactive features, the dashboard stands as a powerful showcase of the transformative potential of data exploration. It invites users to not only analyze but truly experience the depth and richness encapsulated within the Top 2000 Spotify charts. Each click becomes a journey of discovery, unveiling new layers of musical history through a harmonious blend of design principles and data visualization prowess. Users of this dashboard can easily traverse the various insights answering their inquiries as to popularity rankings or characteristic variables that contribute or detract from popularity and genre classification while any given metric is comparative to the overarching timeframe of six decades. Due to the versatility of the dashboard, a user can make a specific inquiry or refine to a select decade, characteristic, song, or genre and glean insight to all informative metrics of the dataset.

V. Discussion

Upon analyzing the data, a clear trend regarding the emergence of musical genres over time was discovered. As we moved closer to the present, there was a noticeable increase in the popularity of electronic-based genres, indicating a shift in current musical preferences. Observing the changes in the characteristic variables of the dataset, listeners have opted for preferences in different attributes over time. In the 1950s, valence, BPM, acousticness, and highly vocal songs seemed to reign at the top charts which are relatable attributes to the most popular genre at that time. The 90s contained the liveliest songs followed by the early 2000s containing vibrant songs with high energy and danceability. Additionally, it was observed that when a new genre is introduced, it has a small initial spike in popularity, but its popularity typically peaks a few years later. This is likely because people need time to adjust to the new sound, and in the music world, criticism often precedes appreciation. Several metrics were found to be correlated with popularity. For example, there is a correlation between popularity and energy, liveliness, and valence. The trend is that if a genre is popular, it is more likely to have high levels of these three characteristics. The analysis revealed a changing musical landscape, with an increase in electronic-based genres, consistent initial peaks preceding later popularity, and a correlation between a genre's popularity and its energetic, lively, and positive qualities. Furthermore, the dashboard goes to show that among the most popular artists Mariah Carey with her contribution of “All I Want for Christmas” is the most popular artist and song in the past six decades closely followed by Lady Gaga with her hit song “Shallow” and Calvin Harris’s song “Summer”.

VI. Future Work

We faced limitations while using Tableau during dataset processing. Tableau emphasizes visually appealing aspects of data analysis, which makes it powerful for creating visualizations. However, it is limited in its efficiency for in-depth analytical tasks.

Enhancing Tableau's capabilities to encompass mathematical operations or programmatic elements for deriving metrics, such as ranges, maximums/minimums, and averages, during visualization creation would significantly contribute to generating more robust visualization products and expediting compilation times. Notably, Tableau's reported limit of close to 100 million data points proved misleading, as even visualizations with as few as 2000 data points exhibited flaws and compromised dataset integrity. Moreover, Tableau parameters and filters are static, containing only one value at a time. This is inflexible, especially when the underlying data set changes. Manual updates to parameters are required to align with the new data. The absence of automated mechanisms for this process adds complexity and hinders adaptability. To propose a new visualization tool optimizing the graphic abilities of Tableau, a software including a coinciding programming script for quick queries of the dataset, adaptability of parameters and a more powerful SQL implementation to reduce runtimes and strain on a computer's RAM would significantly increase the already powerful software.

References

[1] Cleveland, W. S., & McGill, R. (1984). Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. Journal of the American Statistical Association, 79(387), 531–554.*)*